GRIPS Discussion Paper 14-03

Simultaneous Selection of Optimal Bandwidths for the Sharp Regression Discontinuity Estimator

Yoichi Arai
Hidehiko Ichimura

April 2014

National Graduate Institute for Policy Studies
7-22-1 Roppongi, Minato-ku,
Tokyo, Japan 106-8677

Simultaneous Selection of Optimal Bandwidths for the Sharp Regression Discontinuity Estimator*

Yoichi Arai† and Hidehiko Ichimura‡

Abstract

We consider the problem of the bandwidth selection for the sharp regression discontinuity (RD) estimator. The sharp RD estimator requires to estimate two conditional mean functions on the left and the right of the cut-off point nonparametrically. We propose to choose two bandwidths, one for each side for the cut-off point, simultaneously in contrast to common single-bandwidth approaches. We show that allowing distinct bandwidths leads to a nonstandard minimization problem of the asymptotic mean square error. To address this problem, we theoretically define and construct estimators of the asymptotically first-order optimal bandwidths that exploit the second-order bias term. The proposed bandwidths contribute to reduce the mean squared error mainly due to their superior bias performance. A simulation study based on designs motivated by existing empirical literatures exhibits a significant gain of the proposed method under the situations where single-bandwidth approaches can become quite misleading.

Key words: Bandwidth selection, local linear regression, regression discontinuity design

*Earlier versions of this paper were titled "Optimal Bandwidth Selection for Differences of Nonparametric Estimators with an Application to the Sharp Regression Discontinuity Design" and presented at the Japanese Economic Association Spring Meeting, the North American Winter Meeting of the Econometric Society, LSE, UC Berkeley and Yale. Valuable comments were received from seminar participants. We are especially grateful to Yoshihiko Nishiyama, Jack Porter and Jim Powell for many helpful comments. We also thank Jens Ludwig and Douglas Miller for making the data used in Ludwig and Miller (2007) publicly available. Yoko Sakai provided expert research assistance. This research was supported by Grants-in-Aid for Scientific Research No. 22243020 and No. 23330070 from the Japan Society for the Promotion of Science.

†National Graduate Institute for Policy Studies (GRIPS), 7-22-1 Roppongi, Minato-ku, Tokyo 106-8677, Japan; yarai@grips.ac.jp

‡Department of Economics, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-0033, Japan; ichimura@e.u-tokyo.ac.jp

1 Introduction

The regression discontinuity (RD) is a quasi-experimental design to evaluate causal effects, which was introduced by Thistlewaite and Campbell (1960). A large number of empirical applications that exploit the RD design can be found in various areas of economics. See Imbens and Lemieux (2008), van der Klaauw (2008), Lee and Lemieux (2010) and DiNardo and Lee (2011) for an overview and lists of empirical researches.

In the sharp RD design, the treatment status changes when a value of the assignment variable exceeds a known cut-off value and a parameter of interest is the average treatment effect at the cut-off point. Figure 1 illustrates the situation motivated by Ludwig and Miller (2007) where the cut-off value is depicted by a dotted vertical line. The solid line on the left and the dashed line on the right of the cut-off point depict the conditional mean function of the potential outcome for untreated conditional on the assignment variable, denoted by $E(Y(0)|X = x)$, where $Y(0)$ is a potential outcome of untreated and X is an assignment variable. Similarly, the dashed line on the left and the solid line on the right of the cut-off point draw the corresponding function for treated, denoted by $E(Y(1)|X = x)$ where $Y(1)$ is an potential outcome of treated. For both functions, the dashed lines are unobserved. The average treatment effect is given by the difference between the two functions but only at the cut-off point can we estimate the difference under the continuity assumption of both functions. This implies that estimating the treatment effect amounts to estimating two functions at the boundary point. Depending upon assumptions under which we are willing to proceed, an appropriate estimation method changes. One of the most frequently used estimation methods is a nonparametric method using the local linear regression (LLR) because of its superior performance at the boundary.

Given a particular nonparametric estimator, it is well recognized that choosing an appropriate smoothing parameter is a key implementation issue about which various methods have been proposed. In the RD setting, the standard approach in empirical researches is to apply the existing methods of bandwidth choices not necessarily tailored to the RD setting. For example, Ludwig and Miller (2005, 2007)

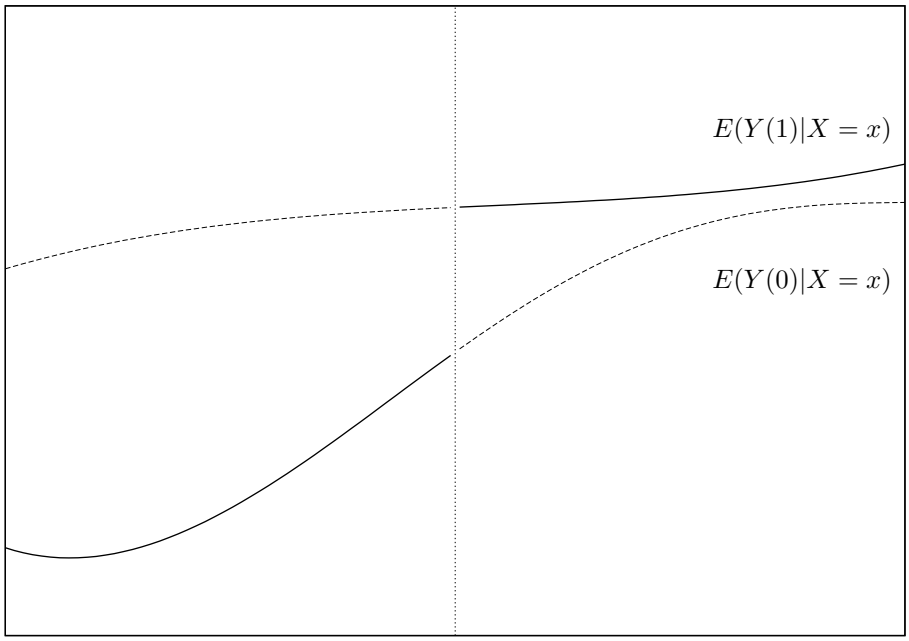


Figure 1. Potential and observed outcomes

(hereafter LM) and DesJardins and McCall (2008) used the cross-validation and the plug-in method, respectively. One notable exception is the bandwidth selection procedure proposed by Imbens and Kalyanaraman (2012) (hereafter IK) to choose the same bandwidth to estimate two functions on both sides of the discontinuity point. The bandwidth proposed by IK is obtained by minimizing the asymptotic approximation of the mean squared error (AMSE) with what they term "regularization".

A single bandwidth approach is familiar to empirical researchers in the applications of matching methods (Abadie and Imbens, 2011) since the supports of covariates for treated and untreated individuals overlap and we wish to construct two comparable groups. This reasoning does not apply to the RD estimator since values of the assignment variable never overlap due to the structure of the RD design. Moreover, the slopes of the conditional mean functions for treated and that for untreated in the vicinity of the cut-off point may be rather different. See Figure 1, for example. A case like this is not an unrealistic artifact and arises naturally in the empirical studies. For example, sharp contrasts in slopes are observed in Figures 1 and 2 in LM, Figures 12 and 14 in DesJardins and McCall (2008), Figures 3 and 5 of Lee (2008) and Figures

1 and 2 of Hinnerich and Pettersson-Lidbom (forthcoming) among others. For the case of Figure 1, considering the bias issue, it would be reasonable to include more of the treated than the untreated because the conditional mean function values vary less for the treated than the untreated. This observation hints at a potential pitfall of common single-bandwidth approaches. We illustrate the usefulness of simultaneously choosing two bandwidths theoretically and through a simulation study based on designs motivated by existing empirical literatures. It exhibits non-negligible gain of choosing distinct bandwidths under the situations where single-bandwidth approaches tend to choose a bandwidth that is too large.

We propose to choose two bandwidths simultaneously based on the AMSE criterion. Although a simultaneous choice of two bandwidths seems natural, it has not yet been considered in the present context.[1] It turns out, this approach leads to a nonstandard problem. We show that when the sign of the product of the second derivatives of the conditional mean functions is negative, the bandwidths that minimize the AMSE are well-defined. But when the sign of the product is positive, the trade-off between bias and variance, which is a key aspect of optimal bandwidth selection, breaks down, and the AMSE can be made arbitrarily small without increasing the bias component. This happens because there exists a specific ratio of bandwidths that can reduce the bias, and we can make the variance arbitrarily small by choosing large values of the bandwidths keeping the ratio constant.

To address this problem, we theoretically define asymptotically first-order optimal (AFO) bandwidths based on objective functions which incorporates a second-order bias term. The AFO bandwidths are defined as the minimizer of the standard AMSE when the sign of the product is negative while they are the minimizer of the AMSE with a second-order bias term subject to the restriction that the first-order bias term is equal to zero when the sign of the product is positive. We show that the AFO bandwidths have advantages over the bandwidths chosen independently regardless of the sign of the product. However the AFO bandwidths are unknown since

[1] Mammen and Park (1997) consider the optimal selection of two bandwidths to estimate the ratio of the first derivative of the density to the density itself. Since the optimal rates for the bandwidths differ in their case, their results do not apply in the present context.

they depend on population quantities. We construct estimators which are shown to be asymptotically equivalent to using the AFO bandwidths. We describe a detailed procedure to implement the proposed method.[2]

We conducted a simulation study to investigate the finite sample properties of the proposed method. Simulation designs are based on the data used in LM and Lee (2008). The first of two main findings is that the performance of the proposed method is robust. The second is that there exists a significant gain in the proposed method under the situations where single-bandwidth approaches tend to choose a bandwidth that is too large. Empirical illustration revisiting the study of LM is also provided.

The paper is organized as follows. We first describe an essential difficulty of the simultaneous selection of the bandwidths and define the AFO bandwidths theoretically to deal with it. We then propose a feasible version of the AFO bandwidth. Finally we illustrate usefulness and practicality via simulation experiments and an empirical example. A detailed procedure for implementation of the proposed method and all proofs are provided in Appendix.

2 Bandwidth Selection of The Sharp Regression Discontinuity Estimators

For individual i we denote potential outcomes by $Y_i(1)$ and $Y_i(0)$, corresponding to outcomes with and without treatment, respectively. Let D_i be a binary variable that stands for the treatment status. Then the observed outcome, Y_i, can be written as $Y_i = D_i Y_i(1) + (1 - D_i)Y_i(0)$. In the sharp RD setting, the treatment status is determined solely by the assignment variable, denoted by X_i: $D_i = \mathbb{I}\{X_i \geq c\}$ where $\mathbb{I}$ denotes the indicator function and c is a known constant. Throughout the paper, we assume that (Y_1, X_1), ..., (Y_n, X_n) are independent and identically distributed observations and X_i has the Lebesgue density f.

Define $m_1(x) = E(Y_i(1)|X_i = x) = E(Y_i|X_i = x)$ for $x \geq c$ and $m_0(x) =$

[2]Matlab and Stata codes to implement the proposed method are available at `http://www3.grips.ac.jp/~yarai/`.

$E(Y_i(0)|X_i = x) = E(Y_i|X_i = x)$ for $x < c$. Suppose that the limits $\lim_{x\to c+} m_1(x)$ and $\lim_{x\to c-} m_0(x)$ exist where $x \to c+$ and $x \to c-$ mean taking the limits from the right and left, respectively. Denote $\lim_{x\to c+} m_1(x)$ and $\lim_{x\to c-} m_0(x)$ by $m_1(c)$ and $m_0(c)$, respectively. Then the average treatment effect at the cut-off point is given by $\tau(c) = m_1(c) - m_0(c)$ and $\tau(c)$ is the parameter of interest in the sharp RD design.[3]

Estimation of $\tau(c)$ requires to estimate two functions, $m_1(c)$ and $m_0(c)$. The nonparametric estimators that we consider are LLR estimators proposed by Stone (1977) and investigated by Fan (1992). For estimating these limits, the LLR is particularly attractive because it exhibits the automatic boundary adaptive property (Fan, 1992, Fan and Gijbels, 1992 and Hahn, Todd, and van der Klaauw, 2001). The LLR estimator for $m_1(c)$ is given by $\hat{\alpha}_{h_1}(c)$, where

$$\left(\hat{\alpha}_{h_1}(c), \hat{\beta}_{h_1}(c)\right) = \arg\min_{\alpha,\beta} \sum_{i=1}^{n} \{Y_i - \alpha - \beta(X_i - c)\}^2 K\left(\frac{X_i - c}{h_1}\right) \mathbb{I}\{X_i \geq c\},$$

where $K(\cdot)$ is a kernel function and h_1 is a bandwidth. A standard choice of the kernel function for the RD estimators is the triangular kernel given by $K(u) = (1-|u|)\mathbb{I}\{|u| < 1\}$ because of its minimax optimality (Cheng, Fan, and Marron, 1997). The solution can be expressed as

$$\begin{bmatrix} \hat{\alpha}_{h_1}(c) \\ \hat{\beta}_{h_1}(c) \end{bmatrix} = \left(X(c)'W_1(c)X(c)\right)^{-1} X(c)'W_1(c)Y,$$

where $X(c)$ is an $n \times 2$ matrix whose ith row is given by $(1, X_i - c)$, $Y = (Y_1, \ldots, Y_n)'$, $W_1(c) = \text{diag}(K_{h_1}(X_i - c))$ and $K_{h_1}(\cdot) = K(\cdot/h_1)\mathbb{I}\{\cdot \geq 0\}/h_1$. The LLR estimator of $m_1(c)$ can also be written as $\hat{\alpha}_{h_1}(c) = e_1' \left(X(c)'W_1(c)X(c)\right)^{-1} X(c)'W_1(c)Y$, where e_1 is a 2×1 vector having one in the first entry and zero in the other entry. Similarly, the LLR estimator for $m_0(c)$, denoted by $\hat{\alpha}_{h_0}(c)$, can be obtained by replacing $W_1(c)$ with $W_0(c)$, where $W_0(c) = \text{diag}(K_{h_0}(X_i - c))$ and $K_{h_0}(\cdot) = K(\cdot/h_0)\mathbb{I}\{\cdot < 0\}/h_0$. Denote $\hat{\alpha}_{h_1}(c)$ and $\hat{\alpha}_{h_0}(c)$ by $\hat{m}_1(c)$ and $\hat{m}_0(c)$, respectively. Then $\tau(c)$ is estimated by $\hat{m}_1(c) - \hat{m}_0(c)$.

[3]See Hahn, Todd, and van der Klaauw (2001).

2.1 The AMSE for The Regression Discontinuity Estimators

In this paper, we propose a simultaneous selection method of two distinct bandwidths, h_1 and h_0, based on an AMSE. This is also the standard approach in the literature.[4]

The conditional MSE of the RD estimators given the assignment variable, X, is defined by

$$MSE_n(h) = E\Big[\big\{ [\hat{m}_1(c) - \hat{m}_0(c)] - [m_1(c) - m_0(c)] \big\}^2 \Big| X \Big].$$

where $X = (X_1, X_2, \ldots, X_n)'$.[5] A standard approach is to obtain the AMSE, ignoring higher-order terms, and to choose the bandwidths that minimize that. To do so, we proceed under the following assumptions. (The integral sign $\int$ refers to an integral over the range $(-\infty, \infty)$ unless stated otherwise.)

ASSUMPTION 1 *$K(\cdot) : \mathbb{R} \to \mathbb{R}$ is a symmetric second-order kernel function that is continuous with compact support; i.e., K satisfies the following: $\int K(u)du = 1$, $\int uK(u)du = 0$, and $\int u^2 K(u)du \neq 0$.*

ASSUMPTION 2 *The positive sequence of bandwidths is such that $h_j \to 0$ and $nh_j \to \infty$ as $n \to \infty$ for $j = 0, 1$.*

Assumptions 1 and 2 are standard in the literature of regression function estimation.

Let $\mathcal{D}$ be an open set in $\mathbb{R}$, k be a nonnegative integer, $\mathcal{C}_k$ be the family of k times continuously differentiable functions on $\mathcal{D}$ and $f^{(k)}(\cdot)$ be the kth derivative of $f(\cdot) \in \mathcal{C}_k$. Let $\mathcal{F}_k(\mathcal{D})$ be the collection of functions f such that $f \in \mathcal{C}_k$ and

$$\left| f^{(k)}(x) - f^{(k)}(y) \right| \leq M_k \left| x - y \right|^{\alpha}, \quad \varepsilon < f(z) < M, \quad x, y, z \in \mathcal{D},$$

for some positive M_k, ε and M such that $0 < \varepsilon < M < \infty$ and some α such that $0 < \alpha \leq 1$.

[4] As IK emphasize, the bandwidth selection problem in the context of the RD setting is how to choose local bandwidths rather than global bandwidths. Thus, bandwidth selection based on either the asymptotic mean "integrated" squared errors or the cross-validation criterion can never be optimal.

[5] Throughout the paper, we use "h" without a subscript to denote a combination of h_1 and h_0; e.g., $MSE_n(h_1, h_0)$ is written as $MSE_n(h)$.

We use $\sigma_1^2(x)$ and $\sigma_0^2(x)$ to denote the conditional variance of Y_i given $X_i = x$ for $x \geq c$ and $x < c$, respectively. Also define $\sigma_1^2(c) = \lim_{x \to c+} \sigma_1^2(x)$, $\sigma_0^2(c) = \lim_{x \to c-} \sigma_0^2(x)$, $m_1^{(2)}(c) = \lim_{x \to c+} m_1^{(2)}(x)$, $m_0^{(2)}(c) = \lim_{x \to c-} m_0^{(2)}(x)$, $m_1^{(3)}(c) = \lim_{x \to c+} m_1^{(3)}(x)$, $m_0^{(3)}(c) = \lim_{x \to c-} m_0^{(3)}(x)$, $\mu_{j,0} = \int_0^\infty u^j K(u) du$ and $\nu_{j,0} = \int_0^\infty u^j K^2(u) du$ for nonnegative integer j.

ASSUMPTION 3 *The density f is an element of $\mathcal{F}_1(\mathcal{D})$ where $\mathcal{D}$ is an open neighborhood of c.*

ASSUMPTION 4 *Let δ be some positive constant. The conditional mean function m_1 and the conditional variance function σ_1^2 are elements of $\mathcal{F}_3(\mathcal{D}_1)$ and $\mathcal{F}_0(\mathcal{D}_1)$, respectively, where $\mathcal{D}_1$ is a one-sided open neighborhood of c, $(c, c+\delta)$, and $m_1(c)$, $m_1^{(2)}(c)$, $m_1^{(3)}(c)$ and $\sigma_1^2(c)$ exist and are bounded. Similarly, m_0 and σ_0^2 are elements of $\mathcal{F}_3(\mathcal{D}_0)$ and $\mathcal{F}_0(\mathcal{D}_0)$, respectively, where $\mathcal{D}_0$ is a one-sided open neighborhood of c, $(c-\delta, c)$, and $m_0(c)$, $m_0^{(2)}(c)$, $m_0^{(3)}(c)$ and $\sigma_0^2(c)$ exist and are bounded.*

Under Assumptions 1, 2, 3 and 4, we can easily generalize the result obtained by Fan and Gijbels (1992) to get,[6]

$$MSE_n(h) = \left\{ \frac{b_1}{2} \left[m_1^{(2)}(c) h_1^2 - m_0^{(2)}(c) h_0^2 \right] \right\}^2 + \frac{v}{nf(c)} \left\{ \frac{\sigma_1^2(c)}{h_1} + \frac{\sigma_0^2(c)}{h_0} \right\} + o\left(h_1^4 + h_1^2 h_0^2 + h_0^4 + \frac{1}{nh_1} + \frac{1}{nh_0} \right), \tag{1}$$

where

$$b_1 = \frac{\mu_{2,0}^2 - \mu_{1,0}\mu_{3,0}}{\mu_{0,0}\mu_{2,0} - \mu_{1,0}^2}, \quad \text{and} \quad v = \frac{\mu_{2,0}^2 \nu_{0,0} - 2\mu_{1,0}\mu_{2,0}\nu_{1,0} + \mu_{1,0}^2 \nu_{2,0}}{(\mu_{0,0}\mu_{2,0} - \mu_{1,0}^2)^2}.$$

This suggests that we choose the bandwidths to minimize the following AMSE:

$$AMSE_n(h) = \left\{ \frac{b_1}{2} \left[m_1^{(2)}(c) h_1^2 - m_0^{(2)}(c) h_0^2 \right] \right\}^2 + \frac{v}{nf(c)} \left\{ \frac{\sigma_1^2(c)}{h_1} + \frac{\sigma_0^2(c)}{h_0} \right\}. \tag{2}$$

[6]The conditions on the first derivative of f and the third derivatives of m_1 and m_0, described in Assumptions 3 and 4, are not necessary to obtain the result (1). They are stated for later use.

However, this procedure may fail. To see why, let h_1, $h_0 \in H$, where $H = (0, \infty)$, and consider the case in which $m_1^{(2)}(c)m_0^{(2)}(c) > 0$. Now choose $h_0 = [m_1^{(2)}(c)/m_0^{(2)}(c)]^{1/2}h_1$. Then, we have

$$AMSE_n(h) = \frac{v}{nh_1 f(c)} \left\{ \sigma_1^2(c) + \sigma_0^2(c) \left[\frac{m_0^{(2)}(c)}{m_1^{(2)}(c)} \right]^{1/2} \right\}.$$

This implies that the bias component can be removed completely from the AMSE by choosing a specific ratio of bandwidths and the AMSE can be made arbitrarily small by choosing a sufficiently large h_1.

One reason for this nonstandard behavior is that the AMSE given in (2) does not account for higher-order terms. If non-removable higher-order terms for the bias component are present, they should punish the act of choosing large values for bandwidths. In what follows, we incorporate a second-order bias term into the AMSE. The next lemma presents the MSE with a second-order bias term by generalizing the higher-order approximation of Fan, Gijbels, Hu, and Huang (1996).[7]

LEMMA 1 *Suppose Assumptions 1–4 hold. Then, it follows that*

$$\begin{aligned} MSE_n(h) = &\left\{ \frac{b_1}{2} \left[m_1^{(2)}(c)h_1^2 - m_0^{(2)}(c)h_0^2 \right] + \left[b_{2,1}(c)h_1^3 - b_{2,0}(c)h_0^3 \right] + o\left(h_1^3 + h_0^3\right) \right\}^2 \\ &+ \frac{v}{nf(c)} \left\{ \frac{\sigma_1^2(c)}{h_1} + \frac{\sigma_0^2(c)}{h_0} \right\} + o\left(\frac{1}{nh_1} + \frac{1}{nh_0} \right), \end{aligned}$$

where

$$\begin{aligned} b_{2,j}(c) &= (-1)^{j+1} \left\{ \xi_1 \left[\frac{m_j^{(2)}(c)}{2} \frac{f^{(1)}(c)}{f(c)} + \frac{m_j^{(3)}(c)}{6} \right] - \xi_2 \frac{m_j^{(2)}(c)}{2} \frac{f^{(1)}(c)}{f(c)} \right\} \\ \xi_1 &= \frac{\mu_{2,0}\mu_{3,0} - \mu_{1,0}\mu_{4,0}}{\mu_{0,0}\mu_{2,0} - \mu_{1,0}^2}, \quad \textit{and} \quad \xi_2 = \frac{\left(\mu_{2,0}^2 - \mu_{1,0}\mu_{3,0}\right)\left(\mu_{0,0}\mu_{3,0} - \mu_{1,0}\mu_{2,0}\right)}{\left(\mu_{0,0}\mu_{2,0} - \mu_{1,0}^2\right)^2}, \end{aligned}$$

for $j = 0, 1$.

In the literature of regression function estimation, it is common to employ local

[7] Fan, Gijbels, Hu, and Huang (1996) show the higher-order approximation of the MSE for interior points of the support of X. Lemma 1 presents the analogous result for a boundary point.

polynomial regression (LPR) of second-order when the conditional mean function is three times continuously differentiable because it is known to reduce bias (see, e.g., Fan, 1992). However, we have several reasons for confining our attention to the LLR. First, as shown later, we can achieve the same bias reduction without employing the LPR when the sign of the product of the second derivatives is positive. When the sign is negative, the existence of the third derivatives becomes unnecessary. Second, even when we use the LPR, we end up with an analogous problem. For example, the first-order bias term is removed by using the LPR, but when the signs of $b_{2,1}(c)$ and $b_{2,0}(c)$ are the same, the second-order bias term can be eliminated by using an appropriate choice of bandwidths.

Given the expression of Lemma 1, one might be tempted to proceed with an AMSE including the second-order bias term:

$$\left\{\frac{b_1}{2}\left[m_1^{(2)}(c)h_1^2 - m_0^{(2)}(c)h_0^2\right] + \left[b_{2,1}(c)h_1^3 - b_{2,0}(c)h_0^3\right]\right\}^2 + \frac{v}{nf(c)}\left\{\frac{\sigma_1^2(c)}{h_1} + \frac{\sigma_0^2(c)}{h_0}\right\} \tag{3}$$

We show that a straightforward minimization of this AMSE does not overcome the problem discussed earlier. That is, the minimization problem is not well-defined when $m_1^{(2)}(c)m_0^{(2)}(c) > 0$. In particular, we show that one can make the order of the bias term $O(h_1^{k+3})$, with k being an arbitrary positive integer, by choosing $h_0^2 = C(h_1, k)h_1^2$ and $C(h_1, k) = C_0 + C_1h_1 + C_2h_1^2 + C_3h_1^3 + \ldots + C_kh_1^k$ for some constants C_0, C_1, $\ldots$, C_k when the sign of the product of the second derivatives is positive. Given that bandwidths are necessarily positive, we must have $C_0 > 0$, although we allow C_1, C_2, $\ldots$, C_k to be negative. For sufficiently large n and for any k, we always have $C(h_1, k) > 0$ given $C_0 > 0$ and we assume this without loss of generality.

To gain insight, consider choosing $C(h_1, 1) = C_0 + C_1h_1$, where $C_0 = m_1^{(2)}(c)/m_0^{(2)}(c)$.

In this case, the sum of the first- and second-order bias terms is

$$\begin{aligned}&\frac{b_1}{2}\left[m_1^{(2)}(c) - C(h_1,1)m_0^{(2)}(c)\right]h_1^2 + \left[b_{2,1}(c) - C(h_1,1)^{3/2}b_{2,0}(c)\right]h_1^3 \\ &\quad = \left\{-\frac{b_1}{2}C_1 m_0^{(2)}(c) + b_{2,1}(c) - C_0^{3/2}b_{2,0}(c)\right\}h_1^3 + O(h_1^4).\end{aligned}$$

By choosing $C_1 = 2\left[b_{2,1}(c) - C_0^{3/2}b_{2,0}(c)\right] \Big/ \left[b_1 m_0^{(2)}(c)\right]$, one can make the order of bias $O(h_1^4)$. Next, consider $C(h_1,2) = C_0 + C_1 h_1 + C_2 h_1^2$, where C_0 and C_1 are as determined above. In this case,

$$\begin{aligned}&\frac{b_1}{2}\left[m_1^{(2)}(c) - C(h_1,2)m_0^{(2)}(c)\right]h_1^2 + \left[b_{2,1}(c) - C(h_1,2)^{3/2}b_{2,0}(c)\right]h_1^3 \\ &\quad = -\left\{b_1 C_2 m_0^{(2)}(c) + 3C_0^{1/2}C_1 b_{2,0}(c)\right\}h_1^4/2 + O(h_1^5).\end{aligned}$$

Hence, by choosing $C_2 = -3C_0^{1/2}C_1 b_{2,0}(c)/[b_1 m_0^{(2)}(c)]$, one can make the order of bias term $O(h_1^5)$. Similar arguments can be formulated for arbitrary k and the discussion above is summarized in the following lemma.

LEMMA 2 *Suppose Assumptions 1–4 hold. Also suppose* $m_1^{(2)}(c)m_0^{(2)}(c) > 0$. *Then there exist a combination of* h_1 *and* h_0 *such that the AMSE including the second-order bias term defined in (3) becomes*

$$\frac{v}{nh_1 f(c)}\left\{\sigma_1^2(c) + \sigma_0^2(c)\left[\frac{m_1^{(2)}(c)}{m_0^{(2)}(c)}\right]^{1/2}\right\} + O\left(h_1^{k+3}\right).$$

for an arbitrary nonnegative integer k.

This implies that one can make the AMSE arbitrarily small by appropriate choices of h_1 and k, leading to non-existence of the optimal solution. It is straightforward to generalize this discussion to the case of the AMSE with higher-order bias terms.

2.2 AFO Bandwidths

We observed that the optimal bandwidths that minimize the AMSE are not well-defined when the sign of the product of the second derivatives is positive. We also noted that simply introducing higher-order bias terms does not help to avoid disappearance of the trade-off. Hence, we propose a new optimality criterion termed "asymptotic first-order optimality".

First, we discuss the case in which $m_1^{(2)}(c)m_0^{(2)}(c) < 0$. Remember that the standard AMSE is given by equation (2). In this situation, the square of the first-order bias term cannot be removed by any choice of the bandwidths and dominates the second-order bias term asymptotically. That is, there is the standard bias-variance trade-off in this case. Hence, it is reasonable to choose the bandwidths that minimize the standard AMSE given in (2).

When $m_1^{(2)}(c)m_0^{(2)}(c) > 0$, by choosing $h_0^2 = C_0 h_1^2$ with $C_0 = m_1^{(2)}(c)/m_0^{(2)}(c)$, the bias component with the second-order term becomes

$$\left[b_{2,1}(c) - C_0^{3/2} b_{2,0}(c)\right] h_1^3 + o\left(h_1^3\right).$$

unless $m_0^{(2)}(c)^3 b_{2,1}(c)^2 = m_1^{(3)}(c)^3 b_{2,0}(c)^2$. With this bias component, there exists a bias-variance trade-off and the bandwidths can be determined. The above discussion is formalized in the following definition and the resulting bandwidths are termed "AFO bandwidths."

DEFINITION 1 *The AFO bandwidths for the RD estimator* minimize the AMSE defined by

$$AMSE_{1n}(h) = \left\{ \frac{b_1}{2} \left[m_1^{(2)}(c)h_1^2 - m_0^{(2)}(c)h_0^2 \right] \right\}^2 + \frac{v}{nf(c)} \left\{ \frac{\sigma_1^2(c)}{h_1} + \frac{\sigma_0^2(c)}{h_0} \right\}.$$

when $m_1^{(2)}(c)m_0^{(2)}(c) < 0$. Their explicit expressions are given by $h_1^* = \theta^* n^{-1/5}$ and

$h_0^* = \lambda^* h_1^*$, where

$$\theta^* = \left\{ \frac{v\sigma_1^2(c)}{b_1^2 f(c) m_1^{(2)}(c) \left[m_1^{(2)}(c) - \lambda^{*2} m_0^{(2)}(c) \right]} \right\}^{1/5} \quad \text{and} \quad \lambda^* = \left\{ -\frac{\sigma_0^2(c) m_1^{(2)}(c)}{\sigma_1^2(c) m_0^{(2)}(c)} \right\}^{1/3}.$$

When $m_1^{(2)}(c) m_0^{(2)}(c) > 0$, *the AFO bandwidths for the RD estimator* minimize the AMSE defined by

$$AMSE_{2n}(h) = \left\{ b_{2,1}(c) h_1^3 - b_{2,0}(c) h_0^3 \right\}^2 + \frac{v}{nf(c)} \left\{ \frac{\sigma_1^2(c)}{h_1} + \frac{\sigma_0^2(c)}{h_0} \right\}$$

subject to the restriction $m_1^{(2)}(c) h_1^2 - m_0^{(2)}(c) h_0^2 = 0$ under the assumption of $m_0^{(2)}(c)^3 b_{2,1}(c)^2 \neq m_1^{(3)}(c)^3 b_{2,0}(c)^2$. Their explicit expressions are given by $h_1^{**} = \theta^{**} n^{-1/7}$ and $h_0^{**} = \lambda^{**} h_1^{**}$, where

$$\theta^{**} = \left\{ \frac{v \left[\sigma_1^2(c) + \sigma_0^2(c)/\lambda^{**} \right]}{6 f(c) \left[b_{2,1}(c) - \lambda^{**3} b_{2,0}(c) \right]^2} \right\}^{1/7} \quad \text{and} \quad \lambda^{**} = \left\{ \frac{m_1^{(2)}(c)}{m_0^{(2)}(c)} \right\}^{1/2}.$$

Definition 1 is stated assuming that the first- and the second-order bias terms do not vanish simultaneously, i.e., $m_0^{(2)}(c)^3 b_{2,1}(c)^2 \neq m_1^{(2)}(c)^3 b_{2,0}(c)^2$.[8]

The proposed bandwidths are called the AFO bandwidths because the $AMSE_{2n}(h)$ is minimized under the restriction that the first-order bias term is removed when the sign is positive. It is worth noting that the order of the optimal bandwidths exhibits dichotomous behavior depending on the sign of the product of the second derivatives. Let h^* and h^{**} denote (h_1^*, h_0^*) and (h_1^{**}, h_0^{**}), respectively. It is easily seen that the orders of $AMSE_{1n}(h^*)$ and $AMSE_{2n}(h^{**})$ are $O_p(n^{-4/5})$ and $O_p(n^{-6/7})$, respectively. This implies that, when the sign is positive, the AFO bandwidths reduce bias with-

[8]Uniqueness of the AFO bandwidths in each case is verified in Arai and Ichimura (2013b). Definition 1 can be generalized to cover the excluded case in a straightforward manner if we are willing to assume the existence of the fourth derivatives. This case corresponds to the situation in which the first- and the second-order bias terms can be removed simultaneously by choosing appropriate bandwidths and the third-order bias term works as a penalty for large bandwidths. Another excluded case in Definition 1 is when $m_1^{(2)}(c) m_0^{(2)}(c) = 0$. It is also possible to extend the idea of the AFO bandwidths when both $m_1^{(2)}(c) = 0$ and $m_0^{(2)}(c) = 0$ hold. This generalization can be carried out by replacing the role of the first- and the second-order bias terms by the second- and the third order bias terms.

out increasing variance and explains why we need not use the LPR even when the third derivatives of $m_1(\cdot)$ and $m_0(\cdot)$ exist. It is also interesting to note that the bias reduction is possible even when the observations on the right of the cut-off point is independent of those on the left. It is the structure of the parameter of interest which is essential for the bias reduction.

The AFO bandwidths has the advantage of the simultaneous selection of bandwidths over the independent selection of the bandwidths. The independent selection chooses the bandwidths on the left and the right of the cut-off optimally for each function without paying attention to the relationship between the two functions. The independently selected bandwidths based on the AMSE criterion are given by

$$\check{h}_1 = \left\{ \frac{v\sigma_1^2(c)}{b_1^2 f(c) \left[m_1^{(2)}(c) \right]^2} \right\}^{1/5} n^{-1/5} \quad \text{and} \quad \check{h}_0 = \left\{ \frac{v\sigma_0^2(c)}{b_1^2 f(c) \left[m_0^{(2)}(c) \right]^2} \right\}^{1/5} n^{-1/5} \tag{4}$$

and the resulting order of the AMSE is $O_p(n^{-4/5})$. The advantage of the simultaneous selection is apparent when $m_1^{(2)}(c)m_0^{(2)}(c) > 0$ since the AFO bandwidths make the order of the AMSE $O_p(n^{-6/7})$. When $m_1^{(2)}(c)m_0^{(2)}(c) < 0$, we note that the AMSE in equation (2) can be written as

$$\begin{aligned} AMSE_n(h) &= \left\{ \frac{b_1}{2} \left[m_1^{(2)}(c)h_1^2 - m_0^{(2)}(c)h_0^2 \right] \right\}^2 + \frac{v}{nf(c)} \left\{ \frac{\sigma_1^2(c)}{h_1} + \frac{\sigma_0^2(c)}{h_0} \right\} \\ &= \left\{ \frac{b_1}{2} \cdot m_1^{(2)}(c)h_1^2 \right\}^2 + \left\{ \frac{b_1}{2} \cdot m_0^{(2)}(c)h_0^2 \right\}^2 - b_1 m_1^{(2)}(c)m_0^{(2)}(c)h_1^2 h_0^2 \\ &\quad + \frac{v}{nf(c)} \left\{ \frac{\sigma_1^2(c)}{h_1} + \frac{\sigma_0^2(c)}{h_0} \right\} \\ &= AMSE_n^1(h) + AMSE_n^0(h) - b_1 m_1^{(2)}(c)m_0^{(2)}(c)h_1^2 h_0^2 \end{aligned} \tag{5}$$

where

$$\begin{aligned} AMSE_n^1(h) &= \left\{ \frac{b_1}{2} \cdot m_1^{(2)}(c)h_1^2 \right\}^2 + \frac{v}{nf(c)} \cdot \frac{\sigma_1^2(c)}{h_1}, \quad \text{and} \\ AMSE_n^0(h) &= \left\{ \frac{b_1}{2} \cdot m_0^{(2)}(c)h_0^2 \right\}^2 + \frac{v}{nf(c)} \cdot \frac{\sigma_0^2(c)}{h_0}. \end{aligned}$$

As shown above, the difference between the objective functions of the AFO bandwidths and the independently selected bandwidths lies solely in the additional bias term. The bandwidths given in equation (4) minimize $AMSE_n^1(h_1)$ and $AMSE_n^0(h_0)$, respectively. The simultaneous selection is superior to the independent selection since the former takes into account the third term of the right hand side of equation (5) which is always positive when $m_1^{(2)}(c)m_0^{(2)}(c) < 0$. This also implies that the advantage of the simultaneous selection would be larger when the third term is larger.

Before we move on, we briefly note that the asymptotically higher-order optimal bandwidths can be proposed in the same manner under a sufficient smoothness condition. For example, the asymptotically second-order optimal (ASO) bandwidths can be constructed when $m_1^{(2)}(c)m_0^{(2)}(c) > 0$ under the assumption that m_1 and m_0 are four times continuously differentiable in the neighborhood of c. However, we do not pursue this direction further in this paper because of implementation difficulty. More detailed discussions are provided in Arai and Ichimura (2013a)

2.3 Feasible Automatic Bandwidth Choice

The AFO bandwidths are clearly not feasible because they depend on unknown quantities related to $f(\cdot)$, m_1, m_0 and, most importantly, on the sign of the product of the second derivatives.

An obvious plug-in version of the AFO bandwidths can be implemented by estimating the second derivatives, $\hat{m}_1^{(2)}(c)$ and $\hat{m}_0^{(2)}(c)$. Depending on the estimated sign of the product, we can construct the plug-in version of the AFO bandwidths provided in Definition 1. We refer to these as "the direct plug-in AFO bandwidths." They are defined by

$$\hat{h}_1^D = \hat{\theta}_1 n^{-1/5}\mathbb{I}\{\hat{m}_1^{(2)}(c)\hat{m}_0^{(2)}(c) < 0\} + \hat{\theta}_2 n^{-1/7}\mathbb{I}\{\hat{m}_1^{(2)}(c)\hat{m}_0^{(2)}(c) \geq 0\},$$
$$\hat{h}_0^D = \hat{\theta}_1\hat{\lambda}_1 n^{-1/5}\mathbb{I}\{\hat{m}_1^{(2)}(c)\hat{m}_0(c) < 0\} + \hat{\theta}_2\hat{\lambda}_2 n^{-1/7}\mathbb{I}\{\hat{m}_1^{(2)}(c)\hat{m}_0^{(2)}(c) \geq 0\},$$

where

$$\hat{\theta}_1 = \left\{ \frac{v\hat{\sigma}_1^2(c)}{b_1^2 \hat{f}(c)\hat{m}_1^{(2)}(c)\left[\hat{m}_1^{(2)}(c) - \hat{\lambda}_1^2 \hat{m}_0^{(2)}(c)\right]} \right\}^{1/5}, \quad \hat{\lambda}_1 = \left\{ -\frac{\hat{\sigma}_0^2(c)\hat{m}_1^{(2)}(c)}{\hat{\sigma}_1^2(c)\hat{m}_0^{(2)}(c)} \right\}^{1/3}, \quad (6)$$

$$\hat{\theta}_2 = \left\{ \frac{v\left[\hat{\sigma}_1^2(c) + \hat{\sigma}_0^2(c)/\hat{\lambda}_2\right]}{6\hat{f}(c)\left[\hat{b}_{2,1}(c) - \hat{\lambda}_2^3 \hat{b}_{2,0}(c)\right]^2} \right\}^{1/7} \quad \text{and} \quad \hat{\lambda}_2 = \left\{ \frac{\hat{m}_1^{(2)}(c)}{\hat{m}_0^{(2)}(c)} \right\}^{1/2}. \quad (7)$$

These bandwidths switch depending on the estimated sign. We can show that the direct plug-in AFO bandwidths are asymptotically as good as the AFO bandwidths in large samples. That is, we can prove that a version of Theorem 1 below also holds for the direct plug-in AFO bandwidths. However, our unreported simulation experiments show a poor performance of the direct plug-in AFO bandwidths under the designs described in Section 3 since they misjudge the rate of the bandwidths whenever the sign is misjudged. Hence we do not pursue the direct plug-in approach further.

Instead, we propose an alternative procedure for choosing bandwidths that switch between two bandwidths more smoothly. To propose feasible bandwidths, we present a modified version of the AMSE (MMSE) defined by

$$MMSE_n(h) = \left\{ \frac{b_1}{2}\left[m_1^{(2)}(c)h_1^2 - m_0^{(2)}(c)h_0^2\right] \right\}^2 + \left\{ b_{2,1}(c)h_1^3 - b_{2,0}(c)h_0^3 \right\}^2 + \frac{v}{nf(x)}\left\{ \frac{\sigma_1^2(x)}{h_1} + \frac{\sigma_0^2(x)}{h_0} \right\}.$$

A notable characteristic of the MMSE is that the bias component is represented by the sum of the squared first- and the second-order bias terms. A key characteristic of the MMSE is that its bias component cannot be made arbitrarily small by any choices of bandwidths even when the sign is positive, unless $m_0^{(2)}(c)^3 b_{2,1}(c)^2 \neq m_1^{(2)}(c)^3 b_{2,0}(c)^2$. Thus, either term can penalize large bandwidths regardless of the sign, in which case, the MMSE preserves the bias-variance trade-off in contrast to the AMSE with the second-order bias term. More precisely, when $m_1^{(2)}(c)m_0^{(2)}(c) < 0$, the square of the first-order bias term serves as the leading penalty and that of the second-order bias

term becomes the second-order penalty. On the other hand, when $m_1^{(2)}(c)m_0^{(2)}(c) > 0$, the square of the second-order bias term works as the penalty and that of the first-order bias term becomes the linear restriction that shows up in the definition of the AFO bandwidths. In fact, the bandwidths that minimize the MMSE are asymptotically equivalent to the AFO bandwidths. This claim can be proved rigorously as a special case of the following theorem.

We propose a feasible bandwidth selection method based on the MMSE. The proposed method for bandwidth selection can be considered as a generalization of the traditional plug-in method (see, e.g., Wand and Jones, 1994, Section 3.6). Consider the following plug-in version of the MMSE denoted by $\widehat{MMSE}$:

$$\widehat{MMSE}_n(h) = \left\{ \frac{b_1}{2} \left[\hat{m}_1^{(2)}(c)h_1^2 - \hat{m}_0^{(2)}(c)h_0^2 \right] \right\}^2 + \left\{ \hat{b}_{2,1}(c)h_1^3 - \hat{b}_{2,0}(c)h_0^3 \right\}^2 + \frac{v}{n\hat{f}(c)} \left\{ \frac{\hat{\sigma}_1^2(c)}{h_1} + \frac{\hat{\sigma}_0^2(c)}{h_0} \right\}, \tag{8}$$

where $\hat{m}_j^{(2)}(c)$, $\hat{b}_{2,j}(c)$, $\hat{\sigma}_j^2(c)$ and $\hat{f}(c)$ are consistent estimators of $m_j^{(2)}(c)$, $b_{2,j}(c)$, $\sigma_j^2(c)$ and $f(x)$ for $j = 0, 1$, respectively. Let $(\hat{h}_1, \hat{h}_0)$ be a combination of bandwidths that minimizes the MMSE given in (8) and $\hat{h}$ denote $(\hat{h}_1, \hat{h}_0)$. In the next theorem, we show that $(\hat{h}_1, \hat{h}_0)$ is asymptotically as good as the AFO bandwidths in the sense of Hall (1983) (see equation (2.2) of Hall, 1983).

THEOREM 1 *Suppose that the conditions stated in Lemma 1 hold. Assume further that* $\hat{m}_j^{(2)}(c)$, $\hat{b}_{2,j}(c)$, $\hat{f}(c)$ *and* $\hat{\sigma}_j^2(c)$ *satisfy* $\hat{m}_j^{(2)}(c) \to m_j^{(2)}(c)$, $\hat{b}_{2,j}(c) \to b_{2,j}(c)$, $\hat{f}(c) \to f(c)$ *and* $\hat{\sigma}_j^2(c) \to \sigma_j^2(c)$ *in probability for* $j = 0, 1$, *respectively. Then, the following hold.*

(i) *When* $m_1^{(2)}(c)m_0^{(2)}(c) < 0$,

$$\frac{\hat{h}_1}{h_1^*} \to 1, \quad \frac{\hat{h}_0}{h_0^*} \to 1, \quad \textit{and} \quad \frac{\widehat{MMSE}_n(\hat{h})}{MSE_n(h^*)} \to 1$$

in probability.

(ii) *When* $m_1^{(2)}(c)m_0^{(2)}(c) > 0$ *and* $m_0^{(2)}(c)^3 b_{2,1}(c)^2 \neq m_1^{(2)}(c)^3 b_{2,0}(c)^2$

$$\frac{\hat{h}_1}{h_1^{**}} \to 1, \quad \frac{\hat{h}_0}{h_0^{**}} \to 1, \quad \textit{and} \quad \frac{\widehat{MMSE}_n(\hat{h})}{MSE_n(h^{**})} \to 1$$

in probability.

The first part of Theorem 1 (i) and (ii) implies that the bandwidths that minimize the MMSE are asymptotically equivalent to the AFO bandwidths regardless of the sign of the product. The second part shows that the minimized value of the plug-in version of the MMSE is asymptotically the same as the MSE evaluated at the AFO bandwidths. These two findings show that the bandwidths that minimize the MMSE possess the desired asymptotic properties. These findings also justify the use of the MMSE as a criterion function. Theorem 1 requires pilot estimates for $m_j^{(2)}(c)$, $b_{2,j}(c)$, $f(c)$ and $\sigma_j^2(c)$ for $j = 0, 1$. A detailed procedure about how to obtain the pilot estimates is given in the next section.

Fan and Gijbels (1996, Section 4.3) points out that replacing constants depending on a kernel function with finite sample approximations can improve finite sample performance. This leads to the following version of the estimated MMSE:

$$\widehat{MMSE}_n^E(h) = \left\{\tilde{b}_{1,1}(c) - \tilde{b}_{1,0}(c)\right\}^2 + \left\{\tilde{b}_{2,1}(c) - \tilde{b}_{2,0}(c)\right\}^2 + \hat{\sigma}_1^2(c)\tilde{v}_1(c) + \hat{\sigma}_0^2(c)\tilde{v}_0(c), \tag{9}$$

where

$$\tilde{b}_{1,j}(c) = \frac{\hat{m}_j^{(2)}(c)}{2} e_1' \tilde{S}_{n,0,j}^{-1} \tilde{c}_{n,2,j},$$

$$\tilde{b}_{2,j}(c) = \left\{ \frac{\hat{m}_j^{(2)}(c)}{2} \cdot \frac{\hat{f}^{(1)}(c)}{\hat{f}(c)} + \frac{\hat{m}_j^{(3)}(c)}{3!} \right\} e_1' \tilde{S}_{n,0,j}^{-1} c_{n,3,j} - \frac{\hat{m}_j^{(2)}(c)}{2} \cdot \frac{\hat{f}^{(1)}(c)}{\hat{f}(c)} e_1' \tilde{S}_{n,0,j}^{-1} S_{n,1,j} \tilde{S}_{n,0,j}^{-1} \tilde{c}_{n,2,j},$$

$$\tilde{v}_j(x) = e_1' S_{n,0,j}^{-1} T_{n,0,j} S_{n,0,j}^{-1} e_1, \quad \tilde{S}_{n,0,j} = S_{n,0,j} - \frac{\hat{f}^{(1)}(c)}{\hat{f}(c)} S_{n1,j}, \quad \tilde{c}_{n,2,j} = c_{n,2,j} - \frac{\hat{f}^{(1)}(c)}{\hat{f}(c)} c_{n,3,j},$$

$$S_{n,k,j} = \begin{bmatrix} s_{n,k,j} & s_{n,k+1,j} \\ s_{n,k+1,j} & s_{n,k+2,j} \end{bmatrix}, \quad T_{n,k,j} = \begin{bmatrix} t_{n,k,j} & t_{n,k+1,j} \\ t_{n,k+1,j} & t_{n,k+2,j} \end{bmatrix}, \quad c_{n,k,j} = \begin{bmatrix} s_{n,k,j} \\ s_{n,k+1,j} \end{bmatrix},$$

$$s_{n,k,j} = \sum_{i=1}^{n} K_{h_j}(X_i - c)(X_i - c)^k, \quad t_{n,k,j} = \sum_{i=1}^{n} K_{h_j}^2(X_i - c)(X_i - c)^k, \tag{10}$$

for $j = 0, 1$. Let $(\hat{h}_1^E, \hat{h}_0^E)$ minimize the MMSE defined by (9), and let $\hat{h}^E$ denote $(\hat{h}_1^E, \hat{h}_0^E)$. Then, the following extension of Theorem 1 holds.

COROLLARY 1 *Suppose that the conditions stated in Theorem 1 hold for each case. Also assume that the second derivative of the density f exists in the neighborhood of x. Then, the results for $\hat{h}_1$, $\hat{h}_0$ and $\widehat{MMSE}_n(\hat{h})$ also hold for $\hat{h}_1^E$, $\hat{h}_0^E$ and $\widehat{MMSE}_n^E(\hat{h}^E)$.*

3 Simulation

To investigate the finite sample performance of the proposed method, we conducted simulation experiments.

3.1 Simulation Designs

The objective of the RDD application is to estimate $\tau(c)$ defined in Section 2. First we consider four designs motivated by the existing empirical studies, LM and Lee (2008). Designs 1–3 are the ones used for simulation experiments in the present context by IK and Calonico, Cattaneo, and Titiunik (2012) (hereafter CCT). Design 4 tries to mimic the situation considered by LM where they investigate the effect of Head Start

assistance on Head Start spending in 1968. This design corresponds to Panel A of Figure II in Ludwig and Miller (2007, pp. 176).[9]

The designs are depicted in Figure 2. For the first two designs, the sign of the product of the second derivatives is negative. The ratio of the second derivative on the right to the one on the left in absolute value is moderate for Design 1, whereas it is rather large for Design 2. For the next two designs, the sign is positive. Design 3 has exactly the same second derivative on both sides, and Design 4 has a relatively large ratio of second derivatives.

For each design, we consider a normally distributed additive error term with mean zero and standard deviation 0.1295. We use data sets of 500 observations and the results are drawn from 10,000 replications. The specification for the assignment variable is exactly the same as that considered by IK.[10]

3.2 Results

The simulation results are presented in Tables 1 and 2. Table 1 reports the results for Designs 1 and 2. The first column explains the design. The second column reports the method used to obtain the bandwidth(s). MMSE refers to the proposed methods based on $\widehat{MMSE}_n(h)$ in equation (8).[11] IK corresponds to the bandwidth denoted by $\hat{h}_{opt}$ in Table 2 of IK.

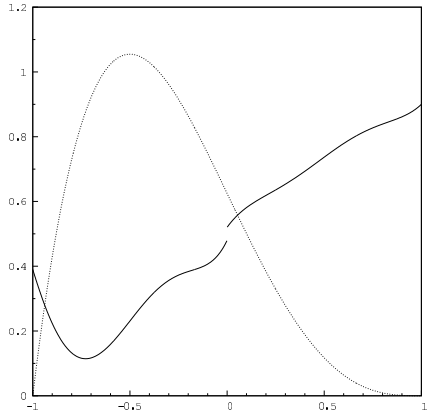
The cross-validation bandwidth used by LM; its implementation is described in Section 4.5 of IK. Note that the cross-validation bandwidth involves one ad hoc parameter although other methods presented here are fully data-driven.[12] DM is the plug-in bandwidths used by DesJardins and McCall (2008) as explained in Section

[9]We followed IK and CCT to obtain the functional form. First we fit the fifth order global polynomial with different coefficients for the right and the left of the cut-off point after rescaling.

[10]In IK the assignment variable is generated by a Beta distribution. More precisely, let Z_i have a Beta distribution with parameters $\alpha = 2$ and $\beta = 4$. Then, the assignment variable X_i is given by $2Z_i - 1$.

[11]As far as the designs considered in this section are concerned, the results based on the methods using $\widehat{MMSE}^E_n(h)$ in equation (9) are almost identical to those using $\widehat{MMSE}_n(h)$. Hence we only show the results based on the latter.

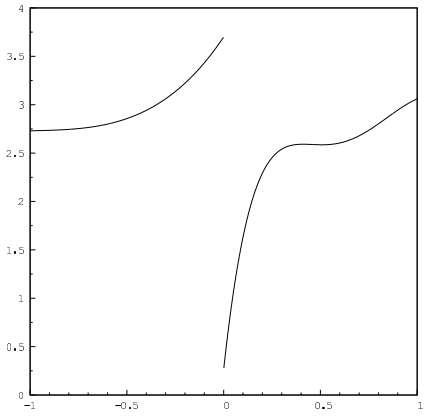
[12]See Section 4.5 of IK for the ad hoc parameter δ used in the cross-validation method. δ is set to 0.5 as in IK.



1. Lee (2008) Data (Design 1 of IK and CCT)

$m_1(x) = 0.52 + 0.84x - 3.0x^2 + 7.99x^3 - 9.01x^4 + 3.56x^5$

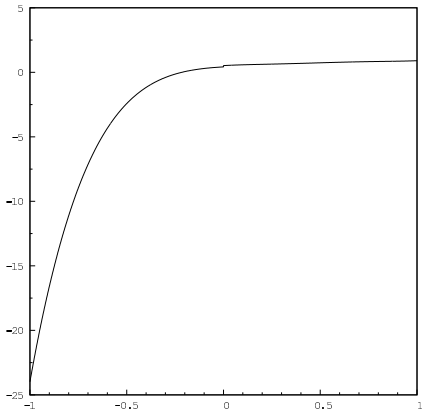
$m_0(x) = 0.48 + 1.27x + 7.18x^2 + 20.21x^3 + 21.54x^4 + 7.33x^5$



2. Ludwig and Miller I (2007) Data (Design 2 of CCT)

$m_1(x) = 0.26 + 18.49x - 54.8x^2 + 74.3x^3 - 45.02x^4 + 9.83x^5$

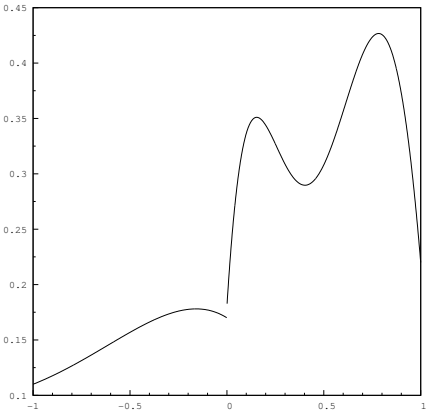
$m_0(x) = 3.70 + 2.99x + 3.28x^2 + 1.45x^3 + 0.22x^4 + 0.03x^5$



3. Constant Additive Treatment Effect (Design 3 of IK)

$m_1(x) = 1.42 + 0.84x - 3.0x^2 + 7.99x^3 - 9.01x^4 + 3.56x^5$

$m_0(x) = 0.42 + 0.84x - 3.0x^2 + 7.99x^3 - 9.01x^4 + 3.56x^5$



4. Ludwig and Miller II (2007, Figure II. B) Data

$m_1(x) = 0.09 + 5.76x - 42.56x^2 + 120.90x^3 - 139.71x^4 + 55.59x^5$

$m_0(z) = 0.03 - 2.26x - 13.14x^2 - 30.89x^3 - 31.98x^4 - 12.1x^5$

Figure 2. Simulation Design (The dotted line in the panel for Design 1 denotes the density of the forcing variable. The supports for $m_1(x)$ and $m_0(x)$ are $x \geq 0$ and $x < 0$, respectively.)

4.4 of IK.[13] DM is an example of the independent bandwidth selection.

The third and fourth columns report the mean (labeled 'Mean') and standard deviation (labeled 'SD') of the bandwidths for IK, LM, and DM. For MMSE, these columns report the bandwidth obtained for the right side of the cut-off point. The fifth and sixth columns report the corresponding ones on the left sides for MMSE. The seventh and eighth columns report the bias (Bias) and the root mean squared error (RMSE) for the sharp RDD estimate, denoted by $\hat{\tau}$. The eighth column report the efficiency relative to the most efficient bandwidth selection rule under Design 1 based on the RMSE.

First, we look at the designs in which the signs of the second derivatives are distinct. The top panel of Table 1, which reports the results for Design 1, demonstrates that all methods perform similarly. DM performs only marginally better. Given similar magnitude for the second derivatives in absolute value, choosing a single bandwidth might be appropriate. The bottom panel of Table 1 reports the results for Design 2, in which there exists a large difference in the magnitudes of the second derivatives. Now MMSE perform significantly better than the other methods, followed by LM. IK and DM perform very poorly mainly because the bandwidths are too large, leading to the large bias. Ignoring the additional bias component represented by the third term in equation (5) is leading to the poor performance of the independence selection (DM). The superior bias performance of MMSE is evident. This shows the importance of choosing a small bandwidth on the right of the cut-off point.

Next, we examine designs in which the sign of the product of the second derivatives is positive. The top panel of Table 2 show that MMSE performs reasonably well for Design 3. The bottom panel of Table 2 reports that MMSE works significantly better than others for Design 4, reflecting the advantage of allowing distinct bandwidths. The bandwidths based on IK, LM and DM tend to be too large for estimating the function on the right of the cut-off and too small on the left relative to the ones based on MMSE.

In summary, for the designs that satisfy the assumptions of Theorem 1, the

[13]The plug-in method used by DesJardins and McCall (2008) is proposed by Fan and Gijbels (1992, 1995).

Table 1: Bias and RMSE for the Sharp RDD, n=500

		$\hat{h}_1$		$\hat{h}_0$		$\hat{\tau}$		
Design	Method	Mean	SD	Mean	SD	Bias	RMSE	Efficiency
Design 1	MMSE	0.378	0.165	0.377	0.151	0.033	0.057	0.895
	IK	0.432	0.114			0.038	0.054	0.944
	LM	0.424	0.118			0.037	0.054	0.944
	DM	0.556	0.135			0.037	0.051	1
Design 2	MMSE	0.076	0.005	0.187	0.027	0.039	0.085	1
	IK	0.177	0.010			0.138	0.151	0.563
	LM	0.129	0.013			0.078	0.107	0.794
	DM	0.267	0.020			0.264	0.272	0.313

Table 2: Bias and RMSE for the Sharp RDD, n=500

		$\hat{h}_1$		$\hat{h}_0$		$\hat{\tau}$		
Design	Method	Mean	SD	Mean	SD	Bias	RMSE	Efficiency
Design 3	MMSE	0.356	0.173	0.205	0.045	-0.021	0.059	0.983
	IK	0.199	0.029			-0.013	0.058	1
	LM	0.112	0.008			-0.003	0.071	0.817
	DM	0.204	0.041			-0.016	0.063	0.921
Design 4	MMSE	0.237	0.094	0.723	0.244	0.025	0.059	1
	IK	0.374	0.127			0.064	0.081	0.728
	LM	0.559	0.205			0.075	0.089	0.663
	DM	0.700	0.264			0.088	0.095	0.621

performance of MMSE is good and stable. IK and DM exhibits disappointing performance for some designs. LM also produces stable results but outperformed by MMSE except Design 1 where LM performs marginally better than MMSE. Overall, MMSE appears very promising.

4 Empirical Illustration

We illustrate how the proposed method in this paper can contribute to empirical researches. In doing so, we revisit the problem considered by LM. They investigate the effect of Head Start on health and schooling. Head Start is the federal government's program aimed to provide preschool, health, and other social services to poor children age three to five and their families. They note that the federal government assisted Head Start proposals of the 300 poorest counties based on the county's 1960 poverty rate and find that the county's 1960 poverty rate can become the assignment variable where the cut-off value is given by 59.1984. They assess the effect of Head Start assistance on numerous measures such as Head Start participation, Head Start spending, other social spending, health, mortality and education.

Here we revisit the study on the effect of Head Start assistance on Head Start spending and mortality provided in Tables II and III of LM. The outcome variables considered in Tables II and III include Head Start spending per child in 1968 and 1972, and the mortality rate for Head Start susceptible causes to all and black children 5 to 9. 1972 Head Start spending per child and the mortality rate for all children generated the simulation Designs 2 and 4 in the previous section, respectively. In obtaining the RD estimates, they employ the LLR using a triangular kernel function as proposed by Porter (2003). For bandwidths, they use 3 different bandwidths, 9, 18 and 36 in somewhat ad-hoc manner rather than relying on some bandwidths selection methods. This implies that the bandwidths and the number of observations with nonzero weight used for estimation are independent of outcome variables.

Table 3 reproduces the results presented in Tables II and III of Ludwig and Miller (2007) for comparison. The point estimates for 1968 Head Start spending per

child range from 114.711 to 137.251 and we might be able to say that they are not very sensitive to the choice of bandwidth. However, the point estimates for 1972 Head Start spending per child range from 88.959 to 182.396. What is more troubling would be the fact that they produce mixed results in statistical significance. For 1968 Head Start spending per child, the point estimate with the bandwidth of 36 produce the result which is statistically significant at 5% level while the estimates with bandwidths of 9 and 18 are not statistically significant even at 10% level. The results for 1972 Head Start spending per child are similar in the sense that the estimates based on the bandwidths of 9 and 36 are statistically significant at 10% level while the estimate based on the bandwidth of 18 is not at the same level.

The results on the mortality rate for all children five to nine exhibit statistical significance though the point estimates range from -1.895 to -1.114 depending on which bandwidth to employ. The point estimate for the mortality rate for black children five to nine with bandwidth 18 is -2.719 which is statistically significant at 5% level while the point estimates with bandwidths 9 and 36 are -2.275 and -1.589, respectively, which are not statistically insignificant even at 10% level. It would be meaningful to see what sophisticated bandwidth selection methods can offer under situations where the results based on ad-hoc approaches cannot be interpreted easily.

Table 4 presents the result based on the bandwidth selection methods based on MMSE and IK. For 1968 Head Start spending per child, the point estimates based on both methods are similar but statistically insignificant although MMSE produces a smaller standard error reflecting the larger bandwidth on the left of the cut-off. The point estimate for 1972 Head Start spending per child differ substantially although they are not statistically significant. For the mortality rate for all children five to nine, both methods produce similar results in terms of the point estimates as well as statistical significance while they generate very different results in both point estimate and statistical significance. To summarize, we found large but statistically insignificant point estimates for Head Start spending and statistically significant estimates for mortality rates by the proposed method in this paper. The results presented in Table 4 alone do not imply any superiority of the proposed method over the existing

Table 3: RD Estimates of the Effect of Head Start Assistance by LM

Variable	Nonparametric		
Bandwidth	9	18	36
Number of observations with nonzero weight	[217, 310]	[287, 674]	[300, 1877]
1968 Head Start spending per child			
RD estimate	137.251	114.711	134.491**
	(128.968)	(91.267)	(62.593)
1972 Head Start spending per child			
RD estimate	182.119*	88.959	130.153*
	(148.321)	(101.697)	(67.613)
Age 5–9, Head Start-related causes, 1973–1983			
RD estimate	−1.895**	−1.198*	−1.114**
	(0.980)	(0.796)	(0.544)
Blacks age 5–9, Head Start-related causes, 1973–1983			
RD estimate	−2.275	−2.719**	−1.589
	(3.758)	(2.163)	(1.706)

This table is reproduced based on Tables II and III of Ludwig and Miller (2007). The numbers of observations with nonzero weight on the right and the left of the cut-off are shown in the square brackets. Standard errors are presented in parentheses. ***, ** and * indicate statistical significance at 1%, 5% and 10% level, respectively.

methods because we never know true causal relationships. However, the results based on the proposed method should provide a meaningful perspective given the simulation experiments demonstrated in the previous section.

5 Conclusion

In this paper, we have proposed a bandwidth selection method for the RD estimators. We provided a discussion on the validity of the simultaneous choice of the bandwidths theoretically and illustrated that the proposed bandwidths can produce good and stable results under situations where single-bandwidth approaches can become misleading based on the simulations motivated by the existing empirical researches.

When we allow two bandwidths to be distinct, we showed that the minimization problem of the AMSE exhibits dichotomous characteristics depending on the sign of the product of the second derivatives of the underlying functions and that the optimal bandwidths that minimize the AMSE are not well-defined when the sign of the product is positive. We introduced the concept of the AFO bandwidths, which

Table 4: RD Estimates of the Effect of Head Start Assistance by MMSE and IK

Variable	MMSE	IK
1968 Head Start spending per child		
Bandwidth	[26.237, 45.925]	19.012
Number of observations with nonzero weight	[299, 2633]	[290, 727]
RD estimate	110.590	108.128
	(76.102)	(80.179)
1972 Head Start spending per child		
Bandwidth	[22.669, 42.943]	20.924
Number of observations with nonzero weight	[298, 2414]	[294, 824]
RD estimate	105.832	89.102
	(79.733)	(84.027)
Age 5–9, Head Start-related causes, 1973–1983		
Bandwidth	[8.038, 14.113]	7.074
Number of observations with nonzero weight	[203, 508]	[182, 243]
RD estimate	-2.094^{***}	-2.359^{***}
	(0.606)	(0.822)
Blacks age 5–9, Head Start-related causes, 1973–1983		
Bandwidth	[22.290, 25.924]	9.832
Number of observations with nonzero weight	[266, 968]	[209, 312]
RD estimate	-2.676^{***}	-1.394
	(1.164)	(2.191)

The bandwidths on the right and the left of the cut-off points are presented in the square brackets. The numbers of observations with nonzero weight on the right and the left of the cut-off are shown in the square brackets. Standard errors are presented in parentheses. ***, ** and * indicate statistical significance based on the bias-corrected t-value at 1%, 5% and 10% level, respectively. See Fan and Gijbels (1996, Section 4.3) for estimation of the bias and variance.

are well-defined regardless of the sign. We proposed a feasible version of the AFO bandwidths. The feasible bandwidths are asymptotically as good as the AFO bandwidths. A simulation study based on designs motivated by existing empirical literatures exhibits non-negligible gain of the proposed method under the situations where a single-bandwidth approach can become quite misleading. We also illustrated the usefulness of the proposed method via an empirical example.

Appendix A Implementation

In this section, we provide a detailed procedure to implement the proposed method in this paper.[14] To obtain the proposed bandwidths, we need pilot estimates of the density, its first derivative, the second and third derivatives of the conditional expectation functions, and the conditional variances at the cut-off point. We obtain these pilot estimates in a number of steps.

Step 1: Obtain pilot estimates for the density $f(c)$ and its first derivative $f^{(1)}(c)$

We calculate the density of the assignment variable at the cut-off point $f(c)$, which is estimated using the kernel density estimator with an Epanechnikov kernel.[15] A pilot bandwidth for kernel density estimation is chosen by using the normal scale rule with Epanechnikov kernel, given by $2.34\hat{\sigma}n^{-1/5}$, where $\hat{\sigma}$ is the square root of the sample variance of X_i (see Silverman, 1986 and Wand and Jones, 1994 for the normal scale rules). The first derivative of the density is estimated by using the method proposed by Jones (1994). The kernel first derivative density estimator is given by $\sum_{i=1}^{n} L((c-X_i)/h)/(nh^2)$, where L is the kernel function proposed by Jones (1994), $L(u) = -15u(1-u^2)1_{\{|u|<1\}}/4$. Again, a pilot bandwidth is obtained by using the normal scale rule, given by $\hat{\sigma} \cdot (112\sqrt{\pi}/n)^{1/7}$.

[14] Matlab and Stata codes to implement the proposed method are available at `http://www3.grips.ac.jp/~yarai/`.

[15] IK estimated the density in a simpler manner (see Section 4.2 of IK). We used the kernel density estimator to be consistent with the estimation method used for the first derivative. Our unreported simulation experiments produced similar results for both methods.

Step 2: Obtain pilot bandwidths for estimating the second and third derivatives $m_j^{(2)}(c)$ and $m_j^{(3)}(c)$ for $j = 0, 1$

We next estimate the second and third derivatives of the conditional mean functions by using the third-order LPR. We obtain pilot bandwidths for the LPR based on the estimated fourth derivatives of $m_1^{(4)}(c) = \lim_{x \to c+} m_1^{(4)}(x)$ and $m_0^{(4)}(c) = \lim_{x \to c-} m_0^{(4)}(x)$. Following IK, we use estimates that are not necessarily consistent by fitting global polynomial regressions. In doing so, we construct a matrix whose ith row is given by $[1 \ (X_i - c) \ (X_i - c)^2 \ (X_i - c)^3 \ (X_i - c)^4]$. This matrix tends to have a high condition number, suggesting potential multicollinearity. That typically makes the polynomial regression estimates very unstable. Hence, we use the ridge regression proposed by Hoerl, Kennard, and Baldwin (1975). This is implemented in two steps. First, using observations for which $X_i \geq c$, we regress Y_i on 1, $(X_i - c)$, $(X_i - c)^2$, $(X_i - c)^3$ and $(X_i - c)^4$ to obtain the standard OLS coefficients $\hat{\gamma}_1$ and the variance estimate $\hat{s}_1^2$. This yields the ridge coefficient proposed by Hoerl, Kennard, and Baldwin (1975): $r_1 = (5\hat{s}_1^2)/(\hat{\gamma}_1'\hat{\gamma}_1)$. Using the data with $X_i < c$, we repeat the procedure to obtain the ridge coefficient, r_0. Let Y be a vector of Y_i, and let X be the matrix whose ith row is given by $[1 \ (X_i - c) \ (X_i - c)^2 \ (X_i - c)^3 \ (X_i - c)^4]$ for observations with $X_i \geq c$, and let I_k be the $k \times k$ identity matrix. The ridge estimator is given by $\hat{\beta}_{r1} = (X'X + r_1 I_5)^{-1} X'Y$, and $\hat{\beta}_{r0}$ is obtained in the same manner. The estimated fourth derivatives are $\hat{m}_1^{(4)}(c) = 24 \cdot \hat{\beta}_{r1}(5)$ and $\hat{m}_0^{(4)}(c) = 24 \cdot \hat{\beta}_{r0}(5)$, where $\hat{\beta}_{r1}(5)$ and $\hat{\beta}_{r0}(5)$ are the fifth elements of $\hat{\beta}_{r1}$ and $\hat{\beta}_{r0}$, respectively. The estimated conditional variance is $\sigma_{r1}^2 = \sum_{i=1}^{n_1} (Y_i - \hat{Y}_i)^2/(n_1 - 5)$, where $\hat{Y}_i$ denotes the fitted values, n_1 is the number of observations for which $X_i \geq c$, and the summation is over i with $X_i \geq c$. σ_{r0}^2 is obtained analogously. The plug-in bandwidths for the third-order LPR used to estimate the second and third derivatives are calculated by

$$h_{\nu,j} = C_{\nu,3}(K) \left(\frac{\sigma_{rj}^2}{\hat{f}(c) \cdot \hat{m}_j^{(4)}(c)^2 \cdot n_j} \right)^{1/9},$$

where $j = 0, 1$ (see Fan and Gijbels, 1996, Section 3.2.3 for information on plug-in bandwidths and the definition of $C_{\nu,3}$). We use $\nu = 2$ and $\nu = 3$ for estimating the second and third derivatives, respectively.

Step 3: Estimation of the second and third derivatives $m_j^{(2)}(c)$ and $m_j^{(3)}(c)$ as well as the conditional variances $\hat{\sigma}_j^2(c)$ for $j = 0, 1$

We estimate the second and third derivatives at the cut-off point by using the third-order LPR with the pilot bandwidths obtained in Step 2. Following IK, we use the uniform kernel, which yields constant values of $C_{2,3} = 5.2088$ and $C_{3,3} = 4.8227$. To estimate $\hat{m}_1^{(2)}(c)$, we construct a vector $Y_a = (Y_1, \ldots, Y_{n_a})'$ and an $n_a \times 4$ matrix, X_a, whose ith row is given by $[1 \ (X_i - c) \ (X_i - c)^2 \ (X_i - c)^3]$ for observations with $c \leq X_i \leq c + h_{2,1}$, where n_a is the number of observations with $c \leq X_i \leq c + h_{2,1}$. The estimated second derivative is given by $\hat{m}_1^{(2)}(c) = 2 \cdot \hat{\beta}_{2,1}(3)$, where $\hat{\beta}_{2,1}(3)$ is the third element of $\hat{\beta}_{2,1}$ and $\hat{\beta}_{2,1} = ({X_a}'X_a)^{-1}X_aY_a$. We estimate $\hat{m}_0^{(2)}(c)$ in the same manner. Replacing $h_{2,1}$ with $h_{3,1}$ leads to an estimated third derivative of $\hat{m}_1^{(3)}(c) = 6 \cdot \hat{\beta}_{3,1}(4)$, where $\hat{\beta}_{3,1}(4)$ is the fourth element of $\hat{\beta}_{3,1}$, $\hat{\beta}_{3,1} = ({X_b}'X_b)^{-1}X_bY_b$, $Y_b = (Y_1, \ldots, Y_{n_b})'$, X_b is an $n_b \times 4$ matrix whose ith row is given by $[1 \ (X_i - c) \ (X_i - c)^2 \ (X_i - c)^3]$ for observations with $c \leq X_i \leq c + h_{3,1}$, and n_b is the number of observations with $c \leq X_i \leq c + h_{3,1}$. The conditional variance at the cut-off point $\sigma_1^2(c)$ is calculated as $\hat{\sigma}_1^2(c) = \sum_{i=1}^{n_2}(Y_i - \hat{Y}_i)^2/(n-4)$, where $\hat{Y}_i$ denotes the fitted values from the regression used to estimate the second derivative.[16] $\hat{\beta}_{2,0}$, $\hat{\beta}_{3,0}$ and $\hat{\sigma}_0^2(c)$ can be obtained analogously.

Step 4: Numerical Optimization

The final step is to plug the pilot estimates into the MMSE given by equation (8) or (9) and to use numerical minimization over the compact region to obtain $\hat{h}_1$ and $\hat{h}_0$. Unlike $AMSE_{1n}(h)$ and $AMSE_{2n}(h)$ subject to the restriction given in Definition 1, the MMSE is not necessarily strictly convex, particularly when the sign of the product is positive. In conducting numerical optimization, it is important to try optimization

[16] One can use the fitted values from the regression used to estimate the third derivatives, having replaced n_a with n_b. These values produce almost identical simulation results.

with several initial values, so as to avoid finding only a local minimum. Either $(\hat{h}_1, \hat{h}_0)$ or $(\hat{h}_1^E, \hat{h}_0^E)$ can be computed as the minimizers depending on the choice of the MMSE.

Appendix B Proofs

Proof of Lemma 1: A contribution to the MSE from a variance component is standard. See Fan and Gijbels (1996) for the details. Here we consider the contribution made by the bias component. We present the proof only for $\hat{\alpha}_{h_1}(c)$. The proof for $\hat{\alpha}_{h_0}$ is parallel and hence is omitted. Denote $\hat{\gamma}_1 = \left(\hat{\alpha}_{h_1}(c),\ \hat{\beta}_{h_1}(c)\right)'$. The conditional bias is given by

$$\text{Bias}(\hat{\gamma}_1|X) = (X(c)'W_1(c)X(c))^{-1}X(c)W_1(c)(m_1 - X(c)\gamma_1),$$

where $m_1 = (m_1(X_1), \ldots, m_1(X_n))'$ and $\gamma_1 = (m_1(c), m_1^{(1)}(c))'$. Note that $S_{n,0,1} = X(c)'W_1(c)X(c)$. The argument made by Fan, Gijbels, Hu, and Huang (1996) can be generalized to yield

$$s_{n,k,1} = nh^k \left\{ f(c)\mu_{k,0} + hf^{(1)}(c)\mu_{k+1,0} + o_p(h) \right\}. \tag{11}$$

Then, it follows that

$$S_{n,0,1} = nH \left\{ f(c)S_{0,1} + hf^{(1)}(c)S_{1,1} + o_p(h) \right\} H,$$

where $H = \text{diag}(1, h)$. By using the fact that $(A + hB)^{-1} = A^{-1} - hA^{-1}BA^{-1} + o(h)$, we obtain

$$S_{n,0,1}^{-1} = n^{-1}H^{-1} \left\{ \frac{1}{f(c)}A_{0,1} - \frac{hf^{(1)}(c)}{f(c)^2}A_{1,1} + o_p(h) \right\} H^{-1}, \tag{12}$$

where

$$A_{0,1} = \begin{bmatrix} \mu_{2,0} & -\mu_{1,0} \\ -\mu_{1,0} & \mu_{0,0}^{-1} \end{bmatrix},$$

$$A_{1,1} = \frac{1}{\mu_{0,0}\mu_{2,0} - \mu_{1,0}^2} \begin{bmatrix} -\mu_{1,0}(\mu_{2,0}^2 - \mu_{1,0}\mu_{3,0}) & \mu_{2,0}(\mu_{2,0}^2 - \mu_{1,0}\mu_{3,0}) \\ \mu_{2,0}(\mu_{2,0}^2 - \mu_{1,0}\mu_{3,0}) & \mu_{1,0}^3 - 2\mu_{0,0}\mu_{1,0}\mu_{2,0} + \mu_{0,0}^2\mu_{3,0} \end{bmatrix}.$$

Next, we consider $X(c)W_1(c)\{m_1 - X(c)\gamma_1\}$. A Taylor expansion of $m_1(\cdot)$ yields

$$X(c)W_1(c)\{m_1 - X(c)\gamma_1\} = \frac{m_1^{(2)}(c)}{2}c_{n,2,1} + \frac{m_1^{(3)}(c)}{3!}c_{n,3,1} + o_p\left(nh^3\right). \tag{13}$$

The definition of $c_{n,k,j}$ in (10), in conjunction with (11), yields

$$c_{n,k,1} = nh^k H\left\{f(c)c_{k,1} + hf^{(1)}(c)c_{k+1,1} + o_p\left(h\right)\right\}. \tag{14}$$

Combining this with (12) and (13) and extracting the first element gives

$$\text{Bias}(\hat{\alpha}_{h_1}(c)|X) = \frac{h^2 b_1 m_1^{(2)}(c)}{2} + b_{2,1}(c)h_1^3 + o_p\left(h_1^3\right).$$

This expression gives the required result. ■

Proof of Theorem 1: Recall that the objective function is:

$$\widehat{MMSE}_n(h) = \left\{\frac{b_1}{2}\left[\hat{m}_1^{(2)}(c)h_1^2 - \hat{m}_0^{(2)}(c)h_0^2\right]\right\}^2 + \left[\hat{b}_{2,1}(c)h_1^3 - \hat{b}_{2,0}(c)h_0^3\right]^2 + \frac{\nu}{n\hat{f}(c)}\left\{\frac{\hat{\sigma}_1^2(c)}{h_1} + \frac{\hat{\sigma}_0^2(c)}{h_0}\right\}.$$

To begin with, we show that $\hat{h}_1$ and $\hat{h}_0$ satisfy Assumption 2. If we choose a sequence of h_1 and h_0 to satisfy Assumption 2, then $\widehat{MMSE}_n(h)$ converges to 0. Assume to the contrary that either one or both of $\hat{h}_1$ and $\hat{h}_0$ do not satisfy Assumption 2. Since $m_0^{(2)}(c)^3 b_{2,1}(c)^2 \neq m_1^{(2)}(c)^3 b_{2,0}(c)^2$ by assumption, $\hat{m}_0^{(2)}(c)^3\hat{b}_{2,1}(c)^2 \neq \hat{m}_1^{(2)}(c)^3\hat{b}_{2,0}(c)^2$ with probability approaching 1. Without loss of generality, we assume

this as well. Then at least one of the first-order bias term, the second-order bias term and the variance term of $\widehat{MMSE}_n(\hat{h})$ does not converge to zero in probability. Then $\widehat{MMSE}_n(\hat{h}) > \widehat{MMSE}_n(h)$ holds for some n. This contradicts the definition of $\hat{h}$. Hence $\hat{h}$ satisfies Assumption 2.

We first consider the case in which $m_1^{(2)}(c)m_0^{(2)}(c) < 0$. In this case, with probability approaching 1, $\hat{m}_1^{(2)}(c)\hat{m}_0^{(2)}(c) < 0$, so that we assume this without loss of generality. When this holds, note that the leading terms are the first term and the last term of $\widehat{MMSE}_n(\hat{h})$ since $\hat{h}_1$ and $\hat{h}_0$ satisfy Assumption 2. Define the plug-in version of $AMSE_{1n}(h)$ provided in Definition 1 by

$$\widehat{AMSE}_{1n}(h) = \left\{ \frac{b_1}{2} \left[\hat{m}_1^{(2)}(c)h_1^2 - \hat{m}_0^{(2)}(c)h_0^2 \right] \right\}^2 + \frac{\nu}{n\hat{f}(c)} \left\{ \frac{\hat{\sigma}_1^2(c)}{h_1} + \frac{\hat{\sigma}_0^2(c)}{h_0} \right\}.$$

A calculation yields $\tilde{h}_1 = \hat{\theta}_1 n^{-1/5} \equiv \tilde{C}_1 n^{-1/5}$ and $\tilde{h}_0 = \hat{\theta}_1 \hat{\lambda}_1 n^{-1/5} \equiv \tilde{C}_0 n^{-1/5}$ where $\hat{\theta}_1$ and $\hat{\lambda}_1$ are defined in (6). With this choice, $\widehat{AMSE}_{1n}(\tilde{h})$ and hence $\widehat{MMSE}_n(\tilde{h})$ converges at the rate of $n^{-4/5}$. Note that if $\hat{h}_1$ or $\hat{h}_0$ converges at the rate slower than $n^{-1/5}$, then the bias term converges at the rate slower than $n^{-4/5}$. If $\hat{h}_1$ or $\hat{h}_0$ converges at the rate faster than $n^{-1/5}$, then the variance term converges at the rate slower than $n^{-4/5}$. Thus the minimizer of $\widehat{MMSE}_n(h)$, $\hat{h}_1$ and $\hat{h}_0$ converges to 0 at rate $n^{-1/5}$.

Thus we can write $\hat{h}_1 = \hat{C}_1 n^{-1/5} + o_p(n^{-1/5})$ and $\hat{h}_0 = \hat{C}_0 n^{-1/5} + o_p(n^{-1/5})$ for some $O_P(1)$ sequences $\hat{C}_1$ and $\hat{C}_0$ that are bounded away from 0 and ∞ as $n \to \infty$. Using this expression,

$$\begin{aligned} \widehat{MMSE}_n(\hat{h}) &= n^{-4/5} \left\{ \frac{b_1}{2} \left[\hat{m}_1^{(2)}(c)\hat{C}_1^2 - \hat{m}_0^{(2)}(c)\hat{C}_0^2 \right] \right\}^2 \\ &\quad + \frac{\nu}{n^{4/5}\hat{f}(c)} \left\{ \frac{\hat{\sigma}_1^2(c)}{\hat{C}_1} + \frac{\hat{\sigma}_0^2(c)}{\hat{C}_0} \right\} + o_p(n^{-4/5}). \end{aligned}$$

Note that

$$\begin{aligned}\widehat{MMSE}_n(\tilde{h}) &= n^{-4/5}\left\{\frac{b_1}{2}\left[\hat{m}_1^{(2)}(c)\tilde{C}_1^2 - \hat{m}_0^{(2)}(c)\tilde{C}_0^2\right]\right\}^2 \\ &\quad + \frac{\nu}{n^{4/5}\hat{f}(c)}\left\{\frac{\hat{\sigma}_1^2(c)}{\tilde{C}_1} + \frac{\hat{\sigma}_0^2(c)}{\tilde{C}_0}\right\} + O_P(n^{-8/5}).\end{aligned}$$

Since $\hat{h}$ is the optimizer, $\widehat{MMSE}_n(\hat{h})/\widehat{MMSE}_n(\tilde{h}) \leq 1$. Thus

$$\frac{\left\{\frac{b_1}{2}\left[\hat{m}_1^{(2)}(c)\hat{C}_1^2 - \hat{m}_0^{(2)}(c)\hat{C}_0^2\right]\right\}^2 + \frac{\nu}{\hat{f}(c)}\left\{\frac{\hat{\sigma}_1^2(c)}{\hat{C}_1} + \frac{\hat{\sigma}_0^2(c)}{\hat{C}_0}\right\} + o_p(1)}{\left\{\frac{b_1}{2}\left[\hat{m}_1^{(2)}(c)\tilde{C}_1^2 - \hat{m}_0^{(2)}(c)\tilde{C}_0^2\right]\right\}^2 + \frac{\nu}{\hat{f}(c)}\left\{\frac{\hat{\sigma}_1^2(c)}{\tilde{C}_1} + \frac{\hat{\sigma}_0^2(c)}{\tilde{C}_0}\right\} + O_P(n^{-4/5})} \leq 1.$$

Note that the denominator converges to

$$\left\{\frac{b_1}{2}\left[m_1^{(2)}(c)C_1^{*2} - m_0^{(2)}(c)C_0^{*2}\right]\right\}^2 + \frac{\nu}{f(c)}\left\{\frac{\sigma_1^2(c)}{C_1^*} + \frac{\sigma_0^2(c)}{C_0^*}\right\},$$

where C_1^* and C_0^* are the unique optimizers of

$$\left\{\frac{b_1}{2}\left[m_1^{(2)}(c)C_1^2 - m_0^{(2)}(c)C_0^2\right]\right\}^2 + \frac{\nu}{f(c)}\left\{\frac{\sigma_1^2(c)}{C_1} + \frac{\sigma_0^2(c)}{C_0}\right\},$$

with respect to C_1 and C_0. This implies that $\hat{C}_1$ and $\hat{C}_0$ also converge to the same respective limit C_1^* and C_0^* because the inequality will be violated otherwise.

Next we consider the case with $m_1^{(2)}(c)m_0^{(2)}(c) > 0$. In this case, with probability approaching 1, $\hat{m}_1^{(2)}(c)\hat{m}_0^{(2)}(c) > 0$, so that we assume this without loss of generality.

When these conditions hold, define $h_0 = \hat{\lambda}_2 h_1$ where $\hat{\lambda}_2$ is defined in (7). This sets the first-order bias term of $\widehat{MMSE}_n(h)$ equal to 0. Define the plug-in version of $AMSE_{2n}(h)$ by

$$\widehat{AMSE}_{2n}(h) = \left\{\hat{b}_{2,1}(c)h_1^3 - \hat{b}_{2,0}(c)h_0^3\right\}^2 + \frac{v}{n\hat{f}(c)}\left\{\frac{\hat{\sigma}_1^2(c)}{h_1} + \frac{\hat{\sigma}_0^2(c)}{h_0}\right\}$$

Choosing h_1 to minimize $\widehat{AMSE}_{2n}(h)$, we define $\tilde{h}_1 = \hat{\theta}_2 n^{-1/7} \equiv \tilde{C}_1 n^{-1/7}$ and $\tilde{h}_0 =$

$\hat{\lambda}_2\tilde{h}_1 \equiv \tilde{C}_0 n^{-1/7}$ where $\hat{\theta}_2$ is defined in (7). Then $\widehat{MMSE}_n(\tilde{h})$ can be written as

$$\widehat{MMSE}_n(\tilde{h}) = n^{-6/7}\left\{\hat{b}_{2,1}(c)\tilde{C}_1^3 - \hat{b}_{2,0}(c)\tilde{C}_0^3\right\}^2 + n^{-6/7}\frac{\nu}{\hat{f}(c)}\left\{\frac{\hat{\sigma}_1^2(c)}{\tilde{C}_1} + \frac{\hat{\sigma}_0^2(c)}{\tilde{C}_0}\right\}.$$

In order to match this rate of convergence, both $\hat{h}_1$ and $\hat{h}_0$ need to converge at the rate slower than or equal to $n^{-1/7}$ because the variance term needs to converge at the rate $n^{-6/7}$ or faster. In order for the first-order bias term to match this rate,

$$\hat{m}_1^{(2)}(c)\hat{h}_1^2 - \hat{m}_0^{(2)}(c)\hat{h}_0^2 \equiv B_{1n} = n^{-3/7}b_{1n},$$

where $b_{1n} = O_P(1)$ so that under the assumption that $m_0^{(2)}(c) \neq 0$, with probability approaching 1, $\hat{m}_0^{(2)}(c)$ is bounded away from 0 so that assuming this without loss of generality, we have $\hat{h}_0^2 = \hat{\lambda}_2^2\hat{h}_1^2 - B_{1n}/\hat{m}_0^{(2)}(c)$. Substituting this expression to the second term and the third term, we have

$$\widehat{MMSE}_n(\hat{h}) = \left\{\frac{b_1}{2}B_{1n}\right\}^2 + \left\{\hat{b}_{2,1}(c)\hat{h}_1^3 - \hat{b}_{2,0}(c)\{\hat{\lambda}_2^2\hat{h}_1^2 - B_{1n}/\hat{m}_0^{(2)}(c)\}^{3/2}\right\}^2 + \frac{\nu}{n\hat{f}(c)}\left\{\frac{\hat{\sigma}_1^2(c)}{\hat{h}_1} + \frac{\hat{\sigma}_0^2(c)}{\{\hat{\lambda}_2^2\hat{h}_1^2 - B_{1n}/\hat{m}_0^{(2)}(c)\}^{1/2}}\right\}.$$

Suppose $\hat{h}_1$ is of order slower than $n^{-1/7}$. Then because $\hat{m}_0^{(2)}(c)^3\hat{b}_{2,1}(c)^2 \neq \hat{m}_1^{(2)}(c)^3\hat{b}_{2,0}(c)^2$ and this holds even in the limit, the second-order bias term is of order slower than $n^{-6/7}$. If $\hat{h}_1$ converges to 0 faster than $n^{-1/7}$, then the variance term converges at the rate slower than $n^{-6/7}$. Therefore we can write $\hat{h}_1 = \hat{C}_1 n^{-1/7} + o_p(n^{-1/7})$ for some $O_P(1)$ sequence $\hat{C}_1$ that is bounded away from 0 and ∞ as $n \to \infty$ and as before $\hat{h}_0^2 = \hat{\lambda}_2^2\hat{h}_1^2 - B_{1n}/\hat{m}_0^{(2)}(c)$. Using this expression, we can write

$$\begin{aligned}\widehat{MMSE}_n(\hat{h}) = {} & n^{-6/7}\left\{\frac{b_1}{2}b_{1n}\right\}^2 \\ & + n^{-6/7}\left\{\left[\hat{b}_{2,1}(c)\hat{C}_1^3 + o_p(1) - \hat{b}_{2,0}(c)\{\hat{\lambda}_2^2\hat{C}_1^2 + o_p(1) - n^{-1/7}b_{1n}/\hat{m}_0^{(2)}(c)\}^{3/2}\right]\right\}^2 \\ & + n^{-6/7}\frac{\nu}{\hat{f}(c)}\left\{\frac{\hat{\sigma}_1^2(c)}{\hat{C}_1 + o_p(1)} + \frac{\hat{\sigma}_0^2(c)}{\{\hat{\lambda}_2^2\hat{C}_1^2 + o_p(1) - n^{-1/7}b_{1n}/\hat{m}_0^{(2)}(c)\}^{1/2}}\right\}.\end{aligned}$$

Thus b_{1n} converges in probability to 0. Otherwise the first-order bias term remains and that contradicts the definition of $\hat{h}_1$.

Since $\hat{h}$ is the optimizer, $\widehat{MMSE}_n(\hat{h})/\widehat{MMSE}_n(\tilde{h}) \leq 1$. Thus

$$\frac{o_p(1) + \left\{\left[\hat{b}_{2,1}(c)\hat{C}_1^3 - \hat{b}_{2,0}(c)\{\hat{\lambda}_2^2\hat{C}_1^2 + o_p(1)\}^{3/2}\right]\right\}^2 + \frac{\nu}{\hat{f}(c)}\left\{\frac{\hat{\sigma}_1^2(c)}{\hat{C}_1+o_p(1)} + \frac{\hat{\sigma}_0^2(c)}{\{\hat{\lambda}_2^2\hat{C}_1^2+o_p(1)\}^{1/2}}\right\}}{\left\{\hat{b}_{2,1}(c)\tilde{C}_1^3 - \hat{b}_{2,0}(c)\tilde{C}_0^3\right\}^2 + \frac{\nu}{\hat{f}(c)}\left\{\frac{\hat{\sigma}_1^2(c)}{\tilde{C}_1} + \frac{\hat{\sigma}_0^2(c)}{\tilde{C}_0}\right\}} \leq 1.$$

If $\hat{C}_1 - \tilde{C}_1$does not converge to 0 in probability, then the ratio is not less than 1 at some point. hence $\hat{C}_1 - \tilde{C}_1 = o_p(1)$. Therefore $\hat{h}_0/\tilde{h}_0$ converges in probability to 1 as well.

The result above also shows that $\widehat{MMSE}_n(\hat{h})/MSE_n(h^*)$ converges to 1 in probability in both cases. ■

Proof of Corollary 1: Observe that equations (11) and (14) imply

$$e_1'\tilde{S}_{n,0,j}^{-1}\tilde{c}_{n,2,j} \to b_1, \quad e_1'\tilde{S}_{n,0,j}^{-1}c_{n,3,j} \to (-1)^{j+1}c_1,$$
$$e_1'\tilde{S}_{n,0,j}^{-1}S_{n,1,j}\tilde{S}_{n,0,j}^{-1}\tilde{c}_{n,2,j} \to (-1)^{j+1}c_2 \quad \text{and} \quad e_1'S_{n,0,j}^{-1}T_{n,0,j}S_{n,0,j}^{-1}e_1 \to v$$

in probability uniformly. With these properties, each step of the proof of Theorem 1 is valid even if $\widehat{MMSE}_n(h)$ is replaced by $\widehat{MMSE}^E{}_n(h)$, thus completing the proof of Corollary 1. ■

References

Abadie, A., and G. W. Imbens (2011): "Bias-corrected matching estimators for average treatment effects," *Journal of Business & Economic Statistics*, 29, 1–11.

Arai, Y., and H. Ichimura (2013a): "Optimal Bandwidth Selection for Differences of Nonparametric Estimators with an Application to the Sharp Regression Discontinuity Design," mimeo.

——— (2013b): "Supplement to Optimal Bandwidth Selection for Differences of

Nonparametric Estimators with an Application to the Sharp Regression Discontinuity Design," mimeo.

Calonico, S., M. D. Cattaneo, and R. Titiunik (2012): "Robust nonparametric bias-corrected inference in the regression discontinuity design," mimeo.

Cheng, M.-Y., J. Fan, and J. S. Marron (1997): "On automatic boundary corrections," *Annals of Statistics*, 25, 1691–1708.

DesJardins, S. L., and B. P. McCall (2008): "The impact of the Gates Millennium scholars program on the retention, college finance- and work-related choices, and future educational aspirations of low-income minority students," mimeo.

DiNardo, J., and D. S. Lee (2011): "Program evaluation and research designs," in *Handook of Lavor Economics*, ed. by O. Ashenfelter, and D. Card, vol. 4A, pp. 463–536. Elsevier B. V.

Fan, J. (1992): "Design-adaptive nonparametric regression," *Journal of the American Statistical Association*, 87, 998–1004.

Fan, J., and I. Gijbels (1992): "Variable bandwidth and local linear regression smoothers," *Annals of Statistics*, 20, 2008–2036.

——— (1995): "Data-driven bandwidth selection in local polynomial fitting: Variable bandwidth and spatial adaption," *Journal of the Royal Statistical Society, Series B*, 57, 371–394.

——— (1996): *Local polynomial modeling and its applications*. Chapman & Hall.

Fan, J., I. Gijbels, T.-C. Hu, and L.-S. Huang (1996): "A study of variable bandwidth selection fro local polynomial regression," *Statistica Sinica*, 6, 113–127.

Hahn, J., P. Todd, and W. van der Klaauw (2001): "Identification and estimation of treatment effects with a regression-discontinuity design," *Econometrica*, 69, 201–209.

Hall, P. (1983): "Large sample optimality of least squares cross-validation in density estimation," *Annals of Statistics*, 11, 1156–1174.

Hinnerich, B. T., and P. Pettersson-Lidbom (forthcoming): "Democracy, redistribution, and political participation: Evidence from Sweden 1919-1938," *Econometrica*.

Hoerl, A. E., R. W. Kennard, and K. F. Baldwin (1975): "Ridge regression: some simulations," *Communications in Statistics, Theory and Methods*, 4, 105–123.

Imbens, G. W., and K. Kalyanaraman (2012): "Optimal bandwidth choice for the regression discontinuity estimator," *Review of Economic Studies*, 79, 933–959.

Imbens, G. W., and T. Lemieux (2008): "Regression discontinuity designs: A guide to practice," *Journal of Econometrics*, 142, 615–635.

Jones, M. C. (1994): "On kernel density derivative estimation," *Communications in Statistics, Theory and Methods*, 23, 2133–2139.

Lee, D. S. (2008): "Randomized experiments from non-random selection in U.S. house elections," *Journal of Econometrics*, 142, 675–697.

Lee, D. S., and T. Lemieux (2010): "Regression discontinuity designs in economics," *Journal of Economic Literature*, 48, 281–355.

Ludwig, J., and D. L. Miller (2005): "Does head start improve children's life changes? Evidence from a regression discontinuity design," NBER Working Paper 11702.

——— (2007): "Does head start improve children's life changes? Evidence from a regression discontinuity design," *Quarterly Journal of Economics*, 122, 159–208.

Mammen, E., and B. U. Park (1997): "Optimal smoothing in adaptive location estimation," *Journal of Statistical Planning and Inference*, 58, 333–348.

Porter, J. (2003): "Estimation in the regression discontinuity model," Mimeo.

Silverman, B. W. (1986): *Density estimation for statistics and data analyasis*. Chapman & Hall, London.

Stone, C. J. (1977): "Consistent nonparametric regression," *Annals of Statistics*, 5, 595–645.

Thistlewaite, D., and D. Campbell (1960): "Regression-discontinuity analysis: An alternative to the ex-post facto experiment," *Journal of Educational Psychology*, 51, 309–317.

van der Klaauw, W. (2008): "Regression-discontinuity analysis: A survey of recent developments in economics," *Labour*, 22, 219–245.

Wand, M. P., and M. C. Jones (1994): *Kernel Smoothing*. Chapman & Hall.